Cite as:
Beckmann, J.F. (2014). The umbrella that is too wide and yet too small: Why Dynamic Testing has still not delivered on the promise that was never made. *Journal of Cognitive Education and Psychology, 13(3), 308-323.*

The umbrella that is too wide and yet too small: Why Dynamic Testing has still not delivered on the promise that was never made

Jens F. Beckmann
Durham University
United Kingdom

Durham University
Leazes Road
Durham, DH1 1TA
United Kingdom
Email: j.beckmann@durham.ac.uk

Abstract

In this article I reflect upon potential reasons for the seemingly persistent impression that Dynamic Testing has not delivered on its promise. Potential reasons are embedded in a paradox. On the one hand validity-related expectations towards dynamic tests seem too broad. This includes fuzziness in defining the diagnostic target constructs, a simplistic quantitative focus on conventional validity indices, and overgeneralised expectations regarding incremental validity. At the same time the focus on Dynamic Testing seems too narrow. By introducing three tests of cognitive flexibility, I exemplify that Dynamic Testing has potential which goes beyond the assessment of learning potential in specific sub-populations. My ambition is to help addressing potential users' misconceptions about Dynamic Testing productively.

Keywords: Dynamic Testing; Incremental Validity; Sensitivity and Specificity; Cognitive Flexibility; Learning ability

The umbrella that is too wide and yet too small: Why Dynamic Testing has still not delivered on the promise that was never made

In this paper[1], I will reflect upon the apparently adamant perception that Dynamic Testing has not delivered on its promise. I will argue that a paradox might have evolved under the umbrella term "Dynamic Testing". It is the paradox of being too broad and too narrow.

I was re-reminded on this rather old issue after I recently moved to the North-East of England to take up a position at Durham University's School of Education. One of the exciting things when starting somewhere new is that one has the chance to talk to colleagues outside and far beyond the more or less close circle of collaborators and researchers who are working in the same area or closely related fields as oneself. Unfortunately, this initial phase of reciprocal interest wears off all too quickly. In one of those conversations around the proverbial water cooler (which in most cases is a tea pot in Northern England) I mentioned Dynamic Testing as one of the areas I have worked in. And there it was: One colleague's reaction – who is a well-respected expert in large-scale assessments in education – was partly expected but also quite disconcerting. His response was something like "Interesting! Yes, of course, I have heard about it. But what a shame that they never really have delivered on their promises." I leave it to you to guess which part was expected and which one was worrying. In the following I would like to reflect upon promises, expectations, and perceptions that underpin the apparently quite persistent sentiment about Dynamic Testing. I would like to thank the editor of this Journal for providing me with the opportunity to do that to an extent that would have gone beyond this Darjeeling moment a few months back.

We have been hearing this rather sobering evaluation of the perceived utility of Dynamic Testing for almost 30 years. The fact that we have been hearing it not only from potential test users (or, rather non-users) but also from colleagues who worked or have been working in the field of Dynamic Testing makes it even more disconcerting.

---

[1] This paper is a slightly extended version of a keynote given at the XIV Conference of the International Association for Cognitive Education and Psychology in Leiden, Netherland.

In this article, I intend to share some deliberations of potential reasons for this problem. I will do this in four parts. First, I will speculate about the nature of the promise that Dynamic Testing has supposedly not being able to keep. Then I will argue that a too broad perspective on the potential of Dynamic Testing may contribute to some unrealistic expectations. In the third part I will explain that perspectives on Dynamic Testing might be too narrow at the same time. Whilst reflecting upon this paradox of being too broad and yet too narrow, I will provide suggestions as to how to take forward our efforts (a) to continue developing useful tests and to successfully prove their usefulness, (b) to foster a more differentiated yet more encompassing perspective on Dynamic Testing, which hopefully will help (c) to address potential users' misconceptions of Dynamic Testing more productively.

Promise, what promise?

The first question to ask is what expectations were nurtured by making what promise that was supposedly not being kept. In the context of testing in general, the central "promise" of a test is in regard to its validity. The question in the context of Dynamic Testing is whether the impression of a broken promise can also be linked to validity-related doubts. In case of an affirmative answer to this question one could argue that the accumulated mass of validity studies, reported either in individual papers or compiled in voluminous readers, can be interpreted as sufficient evidence for refuting the claim of a broken promise. In fact, one would have a hard time finding studies where a test that could rightfully claim the label "dynamic test" failed to predict a criterion of some sort. Of course, this argument is not very convincing for at least three reasons: (1) its arbitrariness in what counts as a dynamic test, (2) its neglect of a potential publication bias, and (3) its lack of specificity regarding to the criterion.

A negative response to the question whether the perception of Dynamic testing not delivering on its promise is validity-related derives from the question, how many test-related validity studies does it take to validate a test concept such as Dynamic Testing[2]? The answer here is "none". According to the prominent definition of validity as the appropriateness, meaningfulness and usefulness of the inferences drawn from test scores (e.g., Cronbach, 1971) we have to acknowledge that validity is not about a test concept and, strictly speaking, not even for a test as such.

[2] Lidz & Elliott (2000) would be a rich source for that.

Although this insight might rather be disorientating at first, it draws our attention to another important aspect. Dynamic Testing as an approach to assessment is anything but homogeneous. Dynamic Testing should be seen as an umbrella term for a wide range of assessment tools that share a set of features[3]. The list of test features often associated with Dynamic Testing includes: provision of feedback, hints, thinking prompts, retries, re-testing after training phases and so forth. The purpose of implementing these kinds of features into a test procedure is to elicit information about test takers' learning potential (Hessels, 2009), learning ability (Guthke, 1982), intellectual change potential (Beckmann, 2001), cognitive modifiability (Tzuriel, 2013), reserve capacity (Kliegl & Baltes, 1987), and others. In this context Dynamic Testing is deemed of particular use for specific target populations such as low socioeconomic status (SES) children, disadvantaged minorities, children with learning difficulties and so forth. Although this list is far from being comprehensive it is already diverse enough to signify the heterogeneity of Dynamic Testing. Its heterogeneity, in particular with regard to the diversity of target constructs, imposes a substantial challenge to a validation of tests that employ features of Dynamic Testing.

So, what is Dynamic Testing, anyway? In reference to Guthke and Wiedl's definition (Guthke & Wiedl, 1996, p. 8; Guthke & Beckmann, 2000, p. 179; Guthke, Beckmann & Wiedl, 2003, p. 225), I see Dynamic Testing as a methodological approach to psychometric assessment that uses systematic variations of task characteristics and / or situational characteristics in the presentation of test items with the intention to evoke intra-individual variability in test performance. Interindividual differences in intraindividual variation are seen as more adequately reflecting the dynamics in the organisation of human behaviour. It is the systematic variation of task and situational characteristics in the item presentation within the test process that justifies the adjective "dynamic" as a qualifier of this assessment approach. Again, may any contention towards this definition that might be experienced by some readers be appreciated as a proof in point regarding the heterogeneity of this test concept. What also becomes apparent is that this definition intentionally lacks any reference to a target construct, i.e. information regarding

[3] Of course, on a different level, one might argue that there is a range of more or less shared epistemological, philosophical, ethical and other factors that underpin the implementation of said features. However, the emphasis would be on "more or less shared" as it is not always apparent that similar things are done for the same reason. Hence, I suggest to "reserve" these considerations for the construct validity discussion.

what it is that we want to assess using Dynamic Testing. This dissociation of means and end, i.e. assessment approach and assessment target, is necessary before we can refocus on the validity issue. In the following, I use the label dynamic tests for assessment procedures that use features of Dynamic Testing as defined.

Too broad …

Test validation often takes the form of looking at correlations between the test scores and some criterion measures, which then is discussed as predictive or criterion validity. In the context of validation of tests that utilise Dynamic Testing there are a range of challenges. Three include the criterion, the construct, and the comparison. A successful prediction of a criterion, which is often interpreted as criterion-related validity, only becomes a constituent of construct validity if the criterion represents an operationalization of the target construct. For example, in a validation of learning tests one would have to expect a substantial link between test scores and a criterion that represents an operationalization of learning ability. The construct relevance of the criterion, of course, needs to be established prior to ascertaining test-criterion correlations to avoid a post hoc reframing of the criterion's relevance that would then help "save the test's reputation"[4]. In the context of alleged under-deliveries on (validity-orientated) promises, where outcomes of dynamic tests do not succeed in predicting some "real-life" criteria such as other test scores, teacher ratings or school grades do not necessarily justify doubts regarding the construct validity of dynamic tests.

The conceptual heterogeneity of Dynamic Testing has consequences with regard to validation strategies. Presuming that "intellectual change potential" does indeed refer to a different construct than, say, "learning ability" and assuming both differ from "cognitive modifiability", "cognitive reserve capacity", or "learning potential" then attempts to establish construct validity for the respective (dynamic) test have to focus on more or less different criteria. As a result, one might be confronted with a situation where, say dynamic tests of learning ability might be more successful than dynamic tests of cognitive modifiability. This, however, would not tell us much about whether Dynamic Testing has or has not delivered on its promise. Each test (that utilises Dynamic Testing) has to demonstrate its value in its own right. In other words, the methodological umbrella of "Dynamic Testing" can

[4] This is where the circular nature of operational definitions of constructs imposes a severe challenge to validation.

provide neither plenipotentiary absolution nor generalised condemnation regarding (validity) promises.

The validity argument cannot be made convincingly by simply referring to sufficiently high correlations between test scores and a (even well-defined and construct-relevant) criterion (see Figure 1a). One might interpret the "promise" of dynamic tests in terms of being able *to better* predict a given criterion than existing assessment tools. This refers to the expectation to provide an increment in information that ultimately will improve " … the appropriateness, meaningfulness and usefulness of the specific inferences made from test scores" (AERA, APA, & NCME, 1995, p. 9; Cronbach, 1971; Messick, 1989). This notion of incremental validity suggests a comparison, which in the case of learning tests, would have to be a comparison to existing or traditional approaches to the assessment of intellectual capacities, in short, intelligence tests. However, a simple comparison of correlation coefficients will not be expedient. Traditional intelligence tests are to a considerable extent a product of a long self-consolidating process in which item and test refinement were informed by the very criteria that are also elements of operational definitions of the construct intelligence itself. Hence, it should not come as a surprise to find a result similar to what is exemplified in Figure 1b, where static, non-dynamic tests are at least as successful in predicting a criterion. Does this mean dynamic tests are not delivering on their promise?

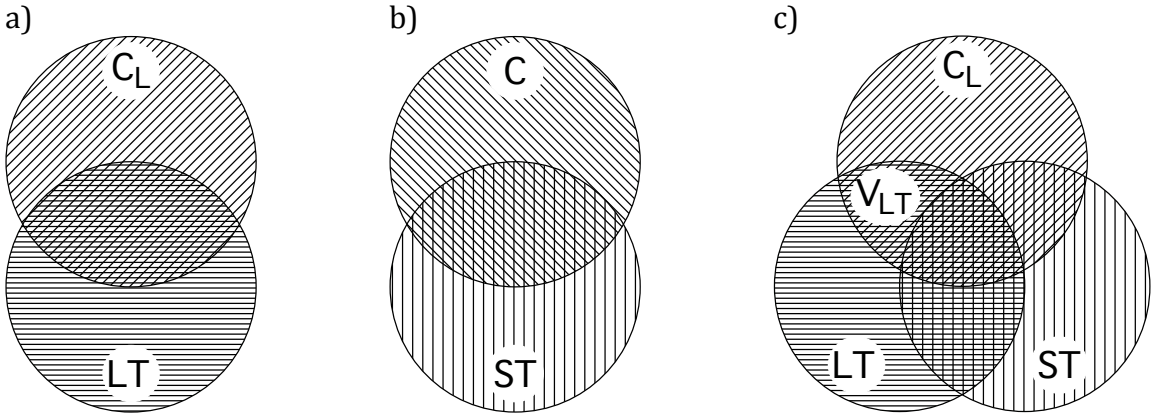


Figure 1: Perspectives on criterion related and construct validation (explanation in text, LT = learning test, ST = static test, C = criterion, C_L = criterion representing an operationalization of learning ability, V_{LT} = incremental validity of LT).

As has been previously argued (e.g., Beckmann & Guthke, 1999, p. 145; Beckmann & Dobat, 2000; Beckmann, 2001, p. 154) the strategy to establish incremental validity of (dynamic) tests needs to build on a construct-related argument, which has a

predominantly *qualitative* focus. In contrast, a *quantitative* comparison of correlation coefficients would merely establish whether a test is successful in measuring more of the same. Learning tests, however, should not be expected to better predict construct-irrelevant criteria. In other words, the quantitative aspect of this comparison is secondary; the raison d'être of dynamic tests rests with their ability to predict behaviours that are qualitatively different from what is measured by traditional non-dynamic (i.e., static) tests. Although traditional static intelligence tests tend to focus on indirect manifestations[5] of learning at best, incremental construct validity of learning tests needs to be established through predictions of *direct* manifestations of the ability to learn, to benefit from feedback, or to respond to mediation etc. In statistical terms, incremental validity refers to LT-specific beta weights in regression analyses as depicted in Figure 1c (see section labelled "V_{LT}"). What also should help resisting the temptation of a simplistic "mine-is-larger-then-yours" competition is the awareness that there are two situations in which section "V_{LT}" could in fact be smaller than its ST-related counterpart (i.e., "V_{ST}") and yet no validity-related pessimism would be warranted. One such situation could result from focussing on an inappropriate criterion. For example a criterion might be an insufficient operationalization of the ability to learn because interindividual differences in performance scores are mainly determined by the amount of prior exposure to learning opportunities. A lack of prediction (i.e., insubstantial "V_{LT}") under these circumstances could cautiously be interpreted as discriminant validity of the learning test under scrutiny.

There is another situation where we in fact expect "V_{LT}" to be smaller than "V_{ST}" to demonstrate construct validity of learning tests (for a more elaborated version of this argument, see Beckmann, 2006). As pointed out earlier, dynamic tests of learning ability seem often to focus on particular sub-populations, such as ethnic minorities, SES-disadvantaged children, or children with learning difficulties. This happens for a reason. The underlying assumption of learning tests is that for those test takers test performance shown under traditional test conditions is not necessarily indicative of their intellectual potential. Therefore, learning opportunities

[5] The prototype of indirect manifestations of the ability to learn is Wechsler's argument that "… the number of words a man knows is at once a measure of his learning ability…" However, for the sake of fairness, it is necessary to add that Wechsler continues in stating that "The one serious stricture that can be made against the vocabulary test as a measure of a man's intelligence is that the number of words a man acquires must necessarily be influenced by his educational and cultural opportunities" (Wechsler, 1935, p. 98-99). This, obviously, represents a plea for more direct measures of learning ability.

are incorporated in the test procedure, which (a) makes tests dynamic and (b) justifies their labelling as learning tests. Test scores, as they are now a construct adequate operationalization of learning ability, are expected to reflect individual differences in the ability to utilise those learning opportunities. Learning opportunities will be offered in form of thinking prompts, hints, or system of graduated feedback after failing to provide a correct answer to a problem presented in the test. Test takers' responsiveness to those mediations represents the operationalization of their potential to learn. For test takers, however, who are able to provide correct answers to test items without an extensive need for mediating support the test procedure becomes more similar to what we find in traditional, non-dynamic tests. For this group of test takers the dynamic tests are not expected to provide incremental information. Figure 2 depicts a situation where panel 2a shows the result pattern expected for the "target population" and panel 2b shows the result pattern likely to be observed in test takers for whom performance scores in non-dynamic, static tests ("ST") represents a sufficiently valid, although indirect estimate of their cognitive potential.

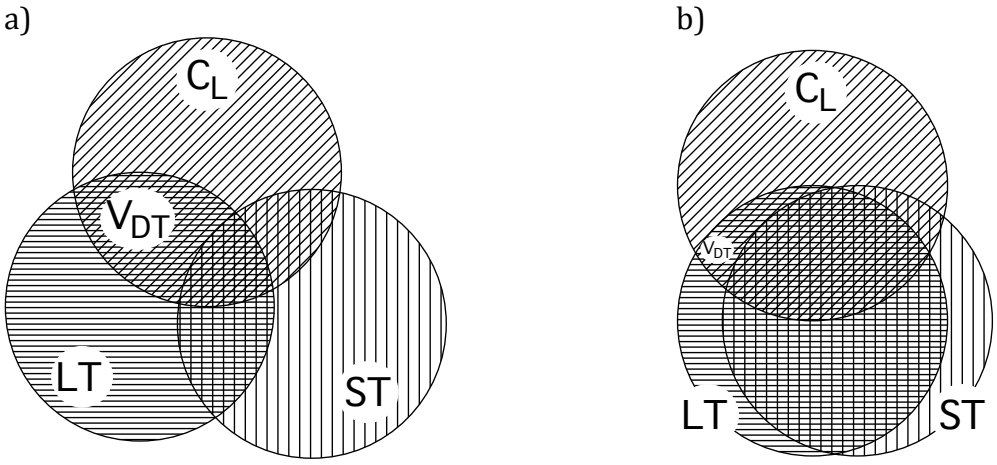


Figure 2: Differential aspect of incremental validity (explanation in text, LT = learning test, ST = static test, C_L = criterion representing an operationalization of learning ability, V_{LT} = incremental validity of LT).

Such result pattern where the same test seems to measure different constructs in different populations indicates so-called *differential validity* (e.g. Urbina, 2004, p. 196), which is usually discussed as a threat to a test's validity. In the context of learning tests, however, I would argue that differential validity is more an indication of construct validity rather than a threat. The major insight that follows from these considerations is that generalised quantitative expectations regarding manifestations

of incremental validity are inappropriate. Validity-related expectations need to be grounded in a priori specifications of construct-relevant behaviour used as a criterion as well as in construct-relevant specifications of a target population. In other words, expectations regarding incremental validity should be constraint to the target population. Otherwise we run the risk of overstating claims (or indirectly encouraging unrealistic expectations) of "superiority always and everywhere" (Beckmann, 2006).

A prominent expectation towards dynamic tests is based on their claim to identify potential rather than manifested abilities. In the context of learning tests, this expectation is nurtured by the assessment process focussing on direct manifestations of learning within the test situation itself. The quality of a diagnostic decision or prediction is dependent on two aspects, sensitivity and specificity. The (doubly) latent nature of potential requires first and foremost high levels of sensitivity in a test for its identification. Sensitivity, in this context, refers to not missing any indications of potential in test takers. Increases in sensitivity can be achieved via a reduction of "false negatives" (Figure 3). To achieve an overall improvement in the quality of prediction (e.g., in form of incremental validity as discussed earlier) increases in sensitivity should not be accompanied by an increase of "false positives", which would constitute a sacrifice of specificity (Beckmann, 2001, p. 156). In the given context, maintaining high levels of specificity means to avoid "seeing" potential where there might be none.

		Criterion:	
		Success	No success
Inference based on test score:	"Potential"	True positive	False positive
	"No potential"	False negative	True negative
		Sensitivity =TP/(TP+FN)	Specificity =TN/(TN+FP)

Figure 3: Sensitivity and Specificity in the context of diagnostic categorisation.

So far I have discussed, admittedly quite cursorily, a few issues that in my view may contribute to unrealistic expectations and subsequently to some rather inauspicious perceptions of Dynamic Testing. The common denominator between these issues seems to be a too broad, or undifferentiated perspective on Dynamic

Testing. A necessary specification and clarification is to be achieved by considering that (a) Dynamic Testing as a test concept cannot be validated as such; (b) tests that utilise dynamic testing need to be validated in their own right; (c) as with any test evaluation, a validation of dynamic tests has to start with a clear definition of the target construct, target population and envisioned purpose of their use; (d) a precise construct definition lays the foundation for the qualitative – rather than a simplistic quantitative – focus on the attempts to establish incremental validity, which means to demonstrate whether the dynamic test under scrutiny is able to predict construct-relevant behaviour (operationalized via an appropriate criterion) that competing traditional approaches cannot, and (e) by using tests of learning potential as an example, incremental validity is to be achieved via an increase in the sensitivity aspect of a prediction without sacrificing its specificity.

… and yet too narrow

In the remainder of this paper I will reflect upon the second part of the paradox. After arguing that a too broad and undifferentiated view on Dynamic Testing might facilitate unrealistic expectations, the point I want to make in the following is that the prevalent discussions around Dynamic Testing are at the same time too narrowly focused on (a) a small yet insufficiently sub-differentiated range of constructs, and (b) particular sub-populations. Both issues may also contribute to impressions that Dynamic Testing has not utilised its potential.

Discussions of Dynamic Testing that identify learning ability, or cognitive modifiability (or any other of its conceptual derivatives) as its exclusive target tend to remind me of a "definition" of football (or soccer, for our transatlantic colleagues). After England's loss of the 14th FIFA World Cup semi-finals to Germany, Gary Lineker supposedly said, "Football is a simple game. Twenty-two men chase a ball for 90 minutes and at the end, the Germans always win". As some will appreciate, the Germans in fact do win occasionally (even in Brazil), others, however, will appreciate that they do not win always. Hence, dynamic testing should not automatically be equated with the assessment of learning ability[6].

In the following I propose that the umbrella term "Dynamic Testing" is far more encompassing and accommodating than it is often perceived. The main underpinning argument for utilising dynamic test procedures for the assessment of

[6] What could be said, however, is that tests that aim to measure learning ability should (a) utilise Dynamic Testing, i.e. should be dynamic tests, and (b) should appropriately be labelled as Learning Tests.

learning ability is that test behaviour observed in traditional intelligence tests has limited representativeness regarding a test taker's ability to learn. This stands in contradiction to the fact that the ability to learn is seen as a core aspect of almost every definition of the construct intelligence. In learning tests, however, the incorporation of learning opportunities into the test situation itself enables test takers to demonstrate their ability to learn; hence their test behaviour represents a more direct proxy of the target construct. In short, learning tests, utilising Dynamic Testing, address the issue of a discrepancy between conceptualisation (i.e., how we define the target construct) and operationalization (i.e., how we measure it). This incongruence of conceptualisation and operationalization is more common than we prefer. I argue that the same is true for another ability construct[7], namely cognitive flexibility.

Cognitive Flexibility as the ability to deal with novelty

When consulting research literature for flexibility-related definitions, we find references to an aptitude for changing lines of thinking (Garaigordobil, 2006); to the ability to shift across concepts and situations (Chi, 1997), to adaptive functioning (Hund & Plumert, 2005), to the ability to change one's mindset (Frensch & Sternberg, 1989), to the ability to adjust to changing demands (Lezak, 1995; Scott, 1962), to an ability to switch modes of response (Kossowska, 1996) etc. The unifying element of those and other definitions seems to be the ability to deal with novelty (Sternberg, 1987). What also becomes apparent is that all these definitions look disturbingly similar to definitions of intelligence. Many definitions of intelligence more or less explicitly refer to the ability to deal with novelty or the ability to adjust to changing demands. As one of many examples, according to William Stern, who is considered the originator of the individual differences perspective in psychology, intelligence " … is a general capacity of an individual consciously to adjust his thinking to new requirements: it is general mental adaptability to new problems and conditions in life" (Stern, 1914, p. 3). Carroll's definition of fluid intelligence as the ability to apply a variety of mental operations to solve novel problems, ones that don't benefit from past learning or experience (Carroll, 1993), provides another example for the conceptual overlap between intelligence and the ability to deal with novelty. According to these and many other definitions, "all things flexible" seem to already

[7] This is not to preclude that the same argument can also be made in relation to non-cognitive constructs in the field of personality assessment (see Guthke, Beckmann, & Wiedl, 2003).

be covered by the construct of (fluid) intelligence. If so, then in emphasising the importance of the construct cognitive flexibility we once again might have fallen for what McNemar called the first cardinal principle of psychological progress: Give new names to old things (McNemar, 1964, p. 872). In my view, however, the alarming familiarity refers rather to the situation described in the context of the measurement of learning ability. As it appears, it is another example of the discrepancy between conceptualisation and operationalisation. Based on the premise that test behaviour should be indicative of the ability construct aimed to measure, the question emerges in the context of cognitive flexibility, to what extent do traditional intelligence tests enable examinees to demonstrate their ability to deal with novelty, to change mind sets, to switch modes of responding, to adjust to changes?

Messick (1989) labels a situation where test scores insufficiently reflect the construct the test claims to measure as construct under-representation. The issue of construct under-representation needs to be addressed – as we did with learning and learning tests – by incorporating challenges to the ability to deal with novelty into the test situation. An operationalization of this ability requires an understanding of what we mean by novelty. Experiences of novelty tend to be accompanied with expressions of surprise. In this context, I find it rather surprising how often we are surprised whilst rather need not to be; at the same time it is surprising how often we are not surprised when we rather ought to be. This conundrum has its origins in the fact that novelty tends to come in three disguises. The first form is *ostensible novelty*. This characterises a situation that appears to be novel and unfamiliar although in fact, the previous experience or already acquired knowledge and skills would enable a person to handle the situation competently. An example for ostensible novelty is the transition from left hand traffic to right hand traffic (or vice versa). The second form is *obscured novelty*, where surface features of a situation induce a sense of familiarity whilst the underlying structure of the situation would require a novel approach. An example for such a situation would be the raising of twins. The third form is not really a disguise as such, however, in the potential presence of the other two it might be as challenging to correctly identify it. In case of *objective or "honest" novelty*, things appear to be unfamiliar and they are in fact novel so that existing experience, already acquired knowledge or skills are not necessarily a sufficient basis for handling this situation. Cognitive Flexibility is the ability to tell these situations apart and to act accordingly. The inability to do that may lead to

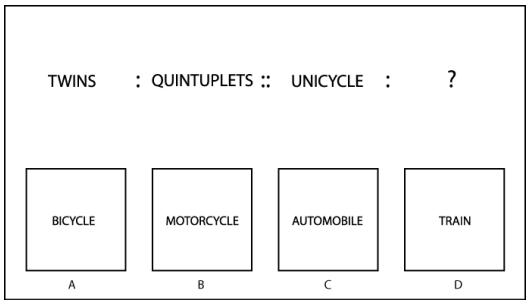
surprising failures. After the brief introduction of the three O's of novelty (i.e., ostensible, obscured and objective), I will now share some ideas of how to implement these into test situations.

Dynamic Testing of cognitive flexibility

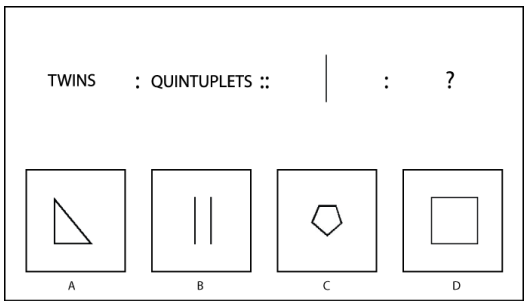
To measure the ability to deal with ostensible novelty we devised the Flexible Mapping Task. The Flexible Mapping Task uses analogies of the type "A is to B as C is to ??". This item paradigm is frequently used in traditional approaches to the assessment of fluid intelligence. The test taker is asked to complete an analogy by selecting the fourth term that in relation to the third term analogously replicates the relationship identified between the first two terms. Figure 4 shows an example of an item using words basis stimuli. For the first item (Figure 4a) the correct answer is AUTOMOBILE because it usually has three wheels more than a unicycle, which replicates the relevant relationship between TWINS and QUINTUPLETS. Traditionally, in analogy items, the relationship identified between the first two elements must be mapped to another elements from the same domain (see Figure 4a). To introduce demands to deal with ostensible novelty we systematically vary the item context and present the same analogy stem again, but now the third term will be from a different domain. For instance, by requiring to map the relationship between two words onto the shape domain (Figure 4b) we expect the application of the same principle (e.g., "three more") in an ostensibly novel situation. We repeat this challenge by now requiring a mapping into the number domain (Figure 4c).

Items in the Flexible Mapping Task are presented in triplets. The first item within each triplet is always a domain-homogenous item (using either words, numbers or shapes as stimuli) whereas the second and third part represents domain-heterogeneous items where mapping in a different domain is required. The intraindividual variability in a test taker's performance across these two item categories (i.e., domain-switching costs) is expected to be indicative of the test taker's ability to deal with ostensible novelty.

a)



b)



c)

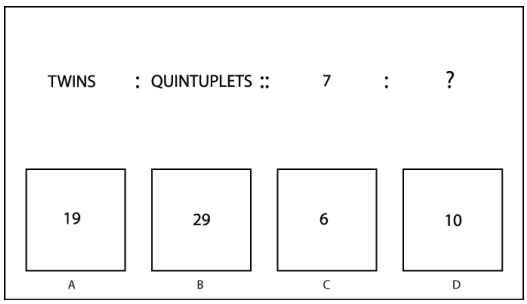


Figure 4: Example item-triplet from Flexible Mapping Task (panel a: domain-homogenous analogy, panels b and c: domain-heterogeneous analogies).

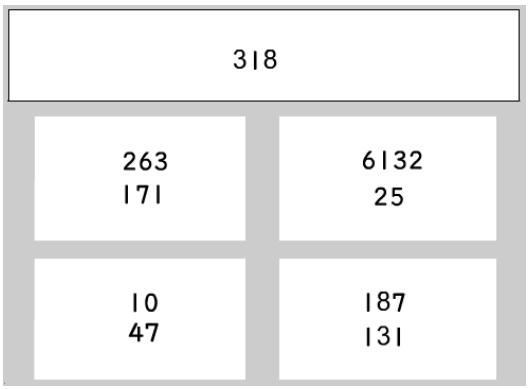
To measure the ability to deal with obscured novelty we propose the Flexible Inference Task. The Flexible Inference Task is a classification task. The test taker is asked to identify the best match to a set target stimulus based on the properties they share. The stimuli in this task are numbers, words or shapes. Figure 5 shows an example of an item using numbers. For the first item (Figure 5a) the correct answer is the upper left pair, because the sum of both numbers equals the target number on top. So far this kind of item does not differ to classification tasks frequently used in traditional tests of fluid intelligence. We introduce the demand to deal with novelty by presenting the same set of stimuli (i.e., same set of numbers), however, rearranged in the subsequent item (Figure 5b). The task remains the same: to find the pair that matches best the number on the top. However, now a successful solution strategy requires a change of the frame of reference. The inferred rule for the

previous problem (arithmetic) is no longer valid. In this example, resistance towards perceiving numbers as stimuli which "invite" the execution of arithmetic operations is required. If numbers are now perceived as graphical patterns that are characterised by features such as angularity or symmetry then it is more likely to find the correct answer to this item (which is the bottom left pair because of the horizontal symmetry of the images). The same set of stimuli is again rearranged into different pairs and the item is presented a third time (Figure 5c). Now an inference based on the number of digits leads to the correct answer (top left corner).

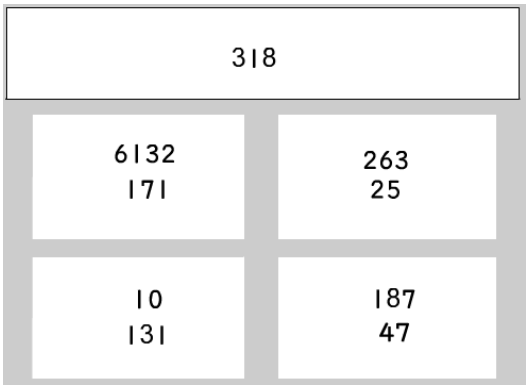
Items in the Flexible Inference Task are presented in triplets (as shown above) and informative feedback as to what the correct solution was and why is given provided after each item. The first items in each triplet require so-called domain-typical inferences (e.g., arithmetic-based with numbers, or semantic meaning with words). The second and third items, in contrast, require domain-atypical inferences (e.g., perceiving numbers as images or number of syllables in words). As an effect, we predict a decrease in performance for the transition from domain-typical to domain-atypical inferences (i.e. switching costs). However, for the transition between the first domain-atypical to the second domain-atypical inference item we in fact expect a slight improvement (i.e. "recovery"). Testing these expectations empirically will be part of the validation strategy (Borsboom, Mellenbergh, & Heerden, 2004). To be successful in the Flexible Inference Task a flexible use of different frames of reference for familiar stimuli is necessary. The ability to inhibit experience gained on previous items is the prerequisite for utilizing different cognitive approaches to the same set of stimuli. Generally, we expected that the intraindividual variability in performance scores caused by the systematic variation within each item triplet will be indicative of test taker's ability to use their cognitive resources flexibly.

To measure the ability to deal with objective novelty we designed the Counterfactual Analogy Task, which is based on an idea discussed by Marr and Sternberg (1986, see also Sternberg & Gastel, 1989). In this task verbal analogies are first presented with preceding statements of familiar facts relating to the analogy stem (factual analogies, see Figure 6a for an example). To introduce the challenge to deal with novelty the same item will then be presented with a preceding counterfactual statement (counterfactual analogy, see Figure 6b).

a)



b)



c)

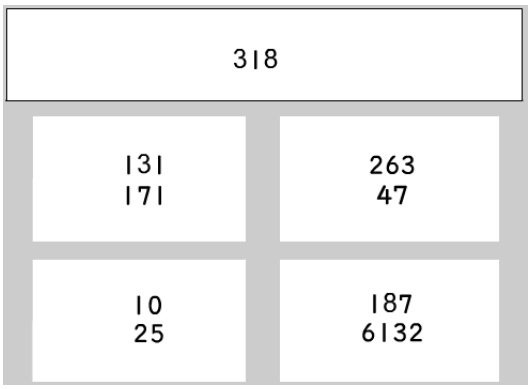
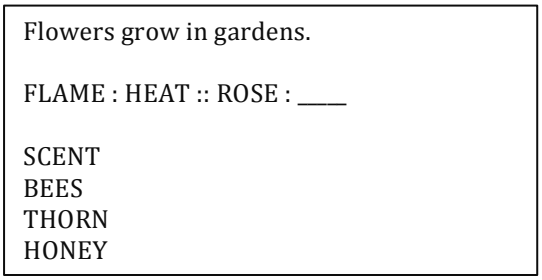


Figure 5: Example item-triplet from Flexible Inference Task (panel a: domain-typical inference, panels b and c: domain-atypical inference)

To solve the counterfactual version of the item pair requires the integration of novel information (i.e., by considering the counterfactual statement to be true) into routine ways of thinking. The intraindividual variability in performance across the item pool comprising both factual and counterfactual analogies is expected to be indicative of the test taker's ability to deal with novelty.

a)



b)

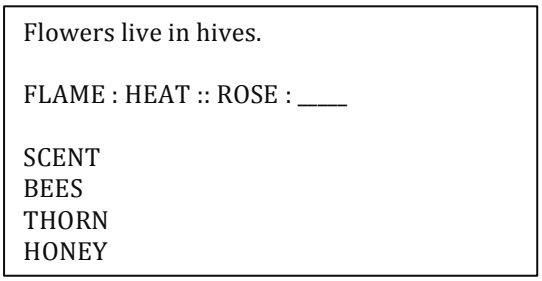


Figure 6: Example item-pair for Counterfactual Verbal Analogies (panel a: factual analogy, panel b: counterfactual analogy).

These three tests aim at measuring the ability to deal with (a) ostensible novelty by requiring to map relationships in analogies flexibly (Flexible Mapping Task), (b) obscured novelty by requiring to flexibly infer relationships in classifications (Flexible Inference Task), and (c) objective novelty by requiring to integrate novel information into routine ways of thinking when solving verbal analogies (Counterfactual Analogies).

These tests are dynamic tests because they use "… systematic variations of task characteristics and / or situational characteristics in the presentation of test items in order to evoke intraindividual variability in test performance" (see definition of Dynamic Testing introduced earlier in this article). The operational focus in these tests is on interindividual differences in intra-individual variation in performance to derive valid estimates of a person's cognitive flexibility.

The test procedure implemented in these flexibility tests is not dissimilar to testing-the-limits, a test approach, when introduced to ability testing (Volle, 1957), attempted to bridge standardised assessment on the one hand and a clinical palpation of performance areas on the other (Boesch, 1964, p. 938; see also Klopfer & Kelley, 1942; Mons, 1955). According to Schmidt (1971, p. 9), in testing-the-limits repeated assessments using the same or parallel tasks under the same or

systematically varying conditions are employed to pursue three main goals: (1) to register intraindividual variability in performance, (2) to determine the modal range of this variability, and (3) to identify the internal as well as external factors which cause this variation. The ultimate goal of testing-the-limits has been to achieve incremental validity to traditional one-off measures. Testing-the-limits can be seen as one of the founding blocks of Dynamic Testing[8]. Other parallels can be seen to so-called load testing in software engineering or cardiac stress tests in health assessments.

For validating these flexibility tests the same principles apply as outlined earlier for dynamic tests aiming at the assessment of learning potential. However, the characteristics of the respective target construct and subsequent target population in combination with the proposed diagnostic purpose of the test determines the qualitative focus in the strategy to establish incremental validity. Whilst the identification of "not-yet-manifested potential" as the diagnostic aim in learning tests translates into an emphasis on increasing sensitivity the strategy to establish incremental validity for flexibility tests has to focus on increasing specificity (Figure 3). In other words, learning tests are expected to find "hidden gems" whereas flexibility tests are expected to find out whether "it is really gold that glitters". Within a Vygotskian framework, I would argue that learning tests aim at the identification of the "zone of proximal development"; cognitive flexibility tests aim at the plasticity of the "zone of current development". One might then speculate that sufficient levels of plasticity in maintaining high levels of (cognitive) performance indicate adequate consolidation of current developmental achievements, which is one precondition for affording a wider horizon when venturing into the zone of proximal development. Addressing these kinds of research questions would be possible with the utilisation of Dynamic Testing.

Neither cognitive flexibility nor any of the learning potential-related constructs are new constructs as such. Conceptually they all are closely linked to intelligence (Guthke & Beckmann, 2001, 2003), where they represent central sub-facets of intellectual functioning. Operationally, however, they are under-represented in traditional approaches to the assessment of intelligence. Dynamic Testing represents a method to redress this validity threat of construct under-representation.

[8] Carlson and Wiedl (1978) give an early example of utilising testing-the limits in a Dynamic Testing procedure.

I used this article as an opportunity to reflect upon potential reasons for the adamant perception that Dynamic Testing has not delivered on its promise. I argue that this perception might be nurtured by a paradox in our and potential users' conception of Dynamic Testing. It is a paradox between being tendentially too broad (or undifferentiated) in our perspective on validity on the one hand and being too narrow in our view on the scope of Dynamic Testing on the other.

In summary, the main points made in the first part of this article are that the only promise Dynamic Testing could possibly make is regarding its usefulness to the psychometric measurement of psychologically relevant constructs. In my view there is no direct way to test whether this promise has been kept. The evaluation of its usability would have to be achieved rather cumulatively via the validation of individual tests that employ Dynamic Testing. This, however, would require that all dynamic tests aim at the same construct, which is clearly not the case (see Lidz & Elliott, 2000 for example). In order to prevent the infamous mixing of apples and oranges[9] validation efforts need to start with a clear definition of the target construct. From that the decision needs to derive what constitutes an appropriate criterion. The construct definition in combination with an explication of the diagnostic purpose of the test also circumscribes the target population, which is instrumental to keeping clearly focused validity expectations. The general nature of the constructs discussed in the context of Dynamic Testing emphasises the importance of incremental validity, which primarily becomes a qualitative matter rather than simply aiming for "more of the same". The construct definition, the proposed diagnostic purpose and the identified target population determine whether incremental validity is to be achieved through an increase in sensitivity or specificity without sacrificing the other. With this line of arguments I wish to insinuate that higher levels of explication and differentiation in our attempts to validate dynamic tests will be beneficial to shaping expectations and to better evaluate which promises are to be kept and which ones should not be made.

In the second part of this paper, I introduced three newly developed tests of cognitive flexibility to demonstrate that Dynamic Testing should not be perceived as being exclusively tied to a particular construct (or construct family). I suspect such

[9] I wonder whether the risk tends to be slightly greater when an orange is called Apfelsine, sinaasappel, or תפוז זהב.

unproductive claims of ownership[10] may contribute to impressions of under-delivery on promises that Dynamic Testing should and, in fact, has never made. Dynamic Testing is nothing more, but certainly nothing less than a methodological approach to assessment. As tools are selected to serve a particular purpose, I optimistically assert that it is our progressed theoretical understanding of abilities as dynamic and malleable phenomena as opposed to static and fixed traits that requires progressive assessment methods to keep pace. I am convinced that Dynamic Testing as a methodological approach to assessment will prove not only beneficial but plain necessary for a valid measurement of psychologically relevant constructs.

[10] This reminds me of one of the bedtime stories my daughter is bound to enjoy these days. In this story, titled "The mushroom in the rain" (Suteyev, V., 1963, *Stories and Pictures*. Moscow: Progress), a range of different animals seeks protection from the pouring rain underneath a mushroom. In the beginning there is only room for an ant, but as the rain continues a butterfly, a mouse, a sparrow, and, yes, even a rabbit finds shelter underneath the parasol mushroom. As I have tried to argue in the second part, the umbrella term "Dynamic Testing" seems broader and more accommodating than it appears. Whilst this view might come with additional challenges, I believe it will help us realise the under-utilised potential of Dynamic Testing. Thus: Let it rain! Well, and it does rain quite often in the North-East of England.

References

Beckmann, J. F. (2001). *Zur Validierung des Konstrukts des intellektuellen Veränderungspotentials* [On the validation of the construct of intellectual change potential]. Berlin: logos.

Beckmann, J. F. (2006). Superiority: Always and everywhere? – On some misconceptions in the validation of Dynamic Testing. *Educational and Child Psychology, 23*, 35-49.

Beckmann, J. F., & Dobat, H. (2000). Zur Validierung der Diagnostik intellektueller Lernfähigkeit. [The validation of the diagnostic of intellectual learning abilty]. *Zeitschrift für Pädagogische Psychologie, 14*, 97-105.

Beckmann, J. F., & Guthke, J. (1999). *Psychodiagnostik des schlußfolgernden Denkens* [The assessment of reasoning ability]. Göttingen: Hogrefe.

Boesch, E. E. (1964). Die diagnostische Systematisierung. [The diagnostic systematization]. In R. Heiss, K. J. Groffmann & L. Michel (Eds.), *Handbuch der Psychologie* (Vol. 6, pp. 930-959). Göttingen: Hogrefe.

Borsboom, D., Mellenbergh, G. J., & Heerden, J. v. (2004). The concept of validity. *Psychological Review, 111*(4), 1061-1071.

Carlson, J. S., & Wiedl, K. H. (1978). Use of testing-the-limits procedures in the assessment of intellectual capabilities in children with learning difficulties. *American Journal of Mental Deficiency, 82*(6), 559-564.

Carroll, J. B. (1993). *Human cognitive abilities – A survey of factoranalytic studies*. New York, NY: Cambridge University Press.

Chi, M. T. (1997). Creativity: Shifting across ontological categories flexibly. In T. B. Ward, S. M. Smith & J. Vaid (Eds.), *Creative thought: An investigation of conceptual structures and processes* (pp. 209-234). Washington DC: American Psychological Association.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational Measurement* (pp. 443-507). Washington, D.C.: American Council on Education.

Frensch, P. A., & Sternberg, R. J. (1989). Expertise and intelligent thinking: When is it worse to know better?

Garaigordobil, M. (2006). Intervention in creativity with children aged 10 and 11 years: Impact of a play program on verbal and graphic-figural creativity. *Creativity Research Journal, 18*(3), 329-345.

Guthke, J. (1982). The learning test concept - An alternative to the traditional static intelligence test. *The German Journal of Psychology, 6*, 306-324.

Guthke, J., & Beckmann, J. F. (2000). Learning test concept and dynamic assessment. In A.

Kozulin & B. Y. Rand (Eds.), *Experience of mediated learning: An impact of Feuerstein's theory in education and psychology* (pp. 175-190). Oxford, UK: Elsevier Science.

Guthke, J., & Beckmann, J. F. (2001). Intelligenz als "Lernfähigkeit"—Lerntests als Alternative zum herkömmlichen Intelligenztest. [Intelligence as the ability to learn – learning tests as alternative to traditional intelligence measures] In E. Stern & J. Guthke (Eds.), *Perspektiven der Intelligenzforschung. Ein Lehrbuch für Fortgeschrittene* (S. 137–161). Lengerich: Pabst.

Guthke, J., & Beckmann, J. F. (2003). Dynamic Assessment with Diagnostic Programs. In R. J. Sternberg & J. Lautrey & T. I. Lubart (Eds.), *Models of intelligence. International perspectives* (pp. 227–242). Washington, DC: APA.

Guthke, J., & Wiedl, K.-H. (1996). *Dynamisches Testen. Zur Psychodiagnostik der intraindividuellen Variabilität* [On the psycho-diagnostic of intraindividual variability]. Göttingen: Hogrefe.

Guthke, J., Beckmann, J. F., & Wiedl, K. H. (2003). Dynamik im Dynamischen Testen [Dynamics in Dynamic Testing]. *Psychologische Rundschau, 54*, 225-232.

Hessels, M. G. P. (2009). Estimation of the predictive validity of the HART by means of a dynamic test of geography. *Journal of Cognitive Education and Psychology, 8*(1), 5-21.

Hund, A. M., & Plumert, J. M. (2005). The stability and flexibility of spatial categories. *Cognitive Psychology, 50,* 1-44.

Kliegl, R. & Baltes, P. B. (1987). Theory-guided analysis of development and aging mechanisms through testing-the limits and research on expertise. In C. Schooler & K. W. Schaie (Eds.), *Cognitive functioning and social structures over the life course*, (pp. 95-119). Norwood: Ablex.

Klopfer, B., & Kelley, D. M. (1942). *The Rorschach technique*. Yonkers, NY: World Book.

Kossowska, M., Matthäus, W., & Necka, E. (1996). The cost of being competent: Expertise and rigidity in coping with novelty. *Polish Psychological Bulletin, 27*(1), 25-38.

Lezak, M. D. (1995). Neuropsychological assessment (3rd ed.). New York, NY: Oxford University Press.

Lidz, C. S., & Elliott, J. G. (Eds.). (2000). *Dynamic assessment: Prevailing models and applications*. Oxford, UK: Elsevier.

Marr, D. B., & Sternberg, R. J. (1986). Analogical reasoning with novel concepts: Differential attention of intellectually gifted and nongifted children to relevant and irrelevant novel stimuli. *Cognitive Development, 1*(1), 53-72.

McNemar, Q., (1964). Lost: Our intelligence? Why? *American Psychologist, 19*, 871-882.

Messick, R. J. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13-103). New York, NY: Macmillan.

Mons, W. E. R. (1955). A normative study of children on the Rorschach test. *Zeitschrift für diagnostische Psychologie und Persönlichkeitsforschung, 3*, 177-180.

Schmidt, L. R. (1971). Testing the limits im Leistungsverhalten: Möglichkeiten und Grenzen [Testing the limits in performance: Potentials and limitations. In E. Duhm (Ed.), Praxis der klinischen Psychologie (Vol. 2, pp. 9-29). Göttingen: Hogrefe.

Scott, W. A. (1962). Cognitive Complexity and Cognitive Flexibility. *Sociometry, 25*(4), 405-414.

Stern, W. (1914). *The psychological methods of testing intelligence* (G. M. Whipple, Trans., German orig. 1912). Baltimore MD: Warwick & York.

Sternberg, R. J. (1987). Coping with novelty and human intelligence. In P. Morris (Ed.), *Modelling cognition* (pp. 57-91). New York, NY: John Wiley & Sons.

Sternberg, R. J., & Gastel, J. (1989). If dancers ate their shoes: Inductive reasoning with factual and counterfactual premises. *Memory & Cognition, 17*(1), 1-10.

Tzuriel, D. (2013). Mediated Learning Experience and Cognitive Modifiability. *Journal of Cognitive Education and Psychology, 12*(1), 59-80.

Urbina, S. (2004). *Essentials of psychological testing*. New York, NY: Wiley.

Volle, F. O. (1957). A proposal for "testing the limits" with mental defectives for the purpose of subtest analysis of the WISC verbal scale. *Journal of Clinical Psychology, 13*, 64-67.

Wechsler, D. (1935). *The range of human capacities*. MD: Williams & Wilkins.